GJCRT : Global Journal of Computer Science Research & Technology
ISSN (online): 2349- 0608

Punjabi Spell Checker

Pushpinder Singh[1] Jaspreet Singh[2]

Assistant Professor, University College Ghanaur, Punjabi University, Patiala, India[1]

Asistant Professor, GSSDGS Khalsa College Patiala, Punjabi University, Patiala, India[2]

pushpinder186@gmail.com[1], jaspreetissaj@gmail.com[2]

ABSTRACT

Spell Checking is one of the applications of Natural Language Processing (NLP). Natural Language Processing is concerned with the interactions between the computer and human languages [2]. Natural language refers to the languages that are spoken by human beings e.g. English, Punjabi, Hindi, French, Spanish etc. Developing spell checkers for human languages is a complex task. Although many spell checkers have been developed by programmers for English and other western languages, but designing a spell checker for Indian languages such as Punjabi poses many new challenges which complicates the design of a spell checker. Punjabi language is different from English and other western language. The use of conjuncts, phonetic nature and character with similar sounds in Punjabi language are the problems that make the design of a spell checker complex. Many existing techniques and algorithms are being used to check the spelling and generate appropriate suggestions for incorrect word of English language. But these algorithms and techniques are not suitable for Punjabi language; rather it needs different algorithms and techniques to develop a spell checker for Punjabi language. The major problem of Punjabi spell checker is that it generates many unrelated suggestions for mis-spelt word. This paper presents the complete design and implementation of **Punjabi Spell Checker.** The objective is to design a Punjabi Spell Checker by checking the Punjabi text with the dictionary of Punjabi words to identify the incorrect words and then to generate the most relevant suggestions out of so many suggestions for incorrect word so that the correct word can be at the top of the suggestions list.

KEYWORDS

Punjabi, Spellchecker, Suggestion List, Typing errors, Ranking of Suggestions, Generation of sugestions.

INTRODUCTION

A spell checker is a computer program which highlights the misspelled word from the document on which the program is being executed. Spell Checker could be a standalone application or part of large software like (Microsoft Office).Spell checkers can be of two types one which only points out the misspelled word and second is which also generate suggestions for the misspelled words along with pointing out the misspelled words and giving options to replace or ignore the specific word [5].

Spelling errors can be partitioned into different categories, that is, real word error and non-word error. A word which is not a valid word is a non-word error. A valid word but not the intended word in the sentence is a real word error [4].

As we can see in Figure 1.1, the working of a spell checker is simple but yet designing the algorithm is the difficult process. The working of spell checker could be understood using the following steps:

1.) First of all, tokenize the input, so that each and every word could be processed.
2.) After tokenizing, word by word repeat the following steps:-
   a) Check the word in the dictionary.
   b) If the word is present in the dictionary that means that word is correct and the program proceeds towards the next word.
      Go to step 2.
   c) If the word is not present in the dictionary, then it means the word might be misspelled. Thus, proceed towards step 3.
3.) Now, the program enters in the error correction mode. In the error correction mode, an algorithm is executed to generate a list of suggestions to replace the misspelled word.
4.) Next, a procedure is executed on the list of generated suggestions to rank the suggestions so that the most appropriate suggestions come on the top.
5.) Now, user can either replace the misspelled word or keep it as it is.
6.) Steps two to five are repeated for each word in the document.

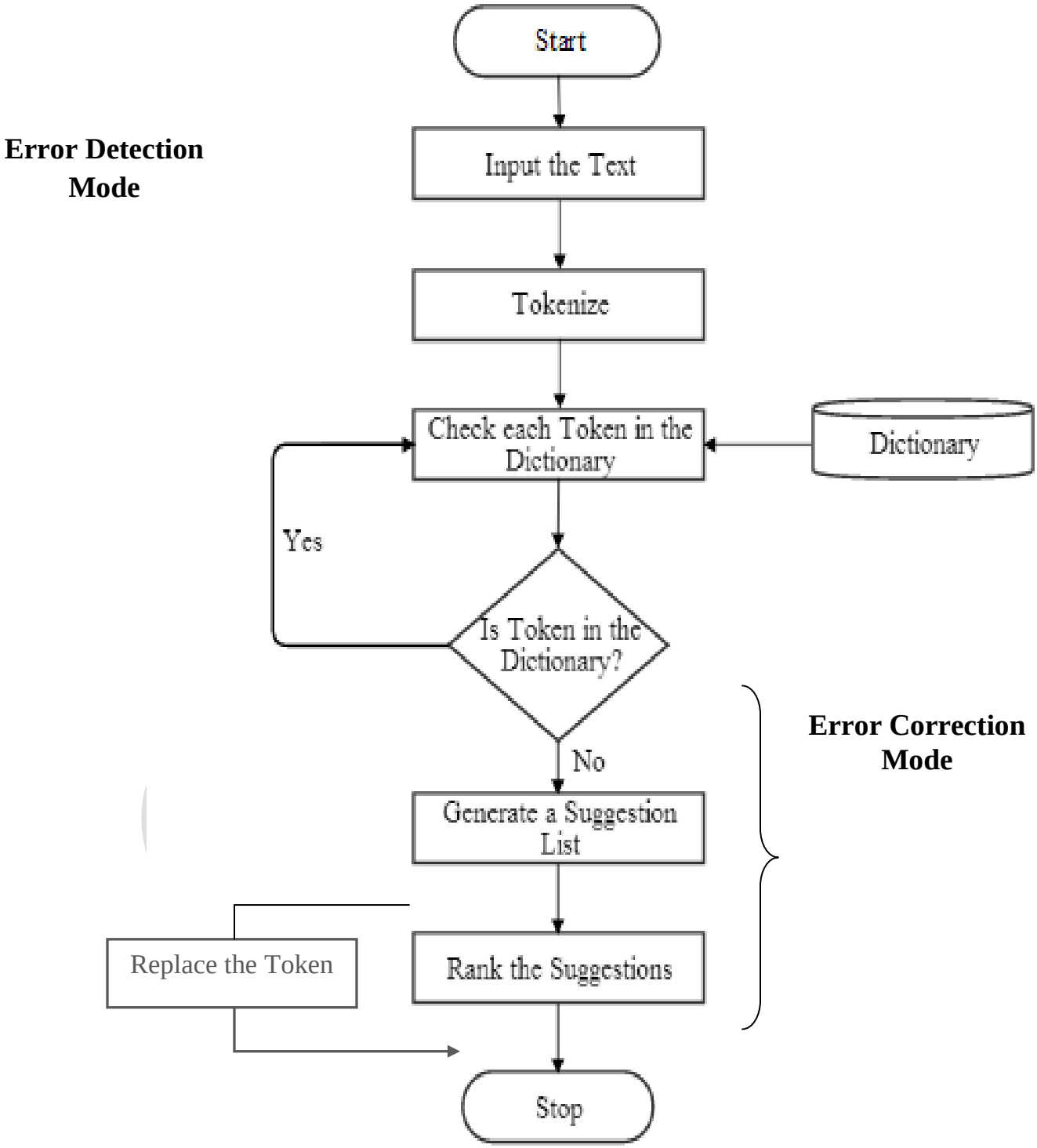

Figure 1.1 Architecture of Spell Checker [6]

Designing a spell checker for Punjabi language is a very complex task as compared to English language. Some of the typical problems faced during designing a spell checker for Punjabi language are [1]:

- There are about forty fonts and keyboard layouts of Punjabi language which are commonly used. And internally, these forty fonts are stored in forty different ways.
- Punjabi language is not written in a linear fashion. The same word can be stored internally in a different way.
- Some of the Punjabi characters have zero width. By mistake, user can make multiple entries of such characters but only single entry is visible.
- There is no perfect word boundary of Punjabi words. One character can be used to show the end of the word, same character can be used as a part of the word.
- There is not standardization of spellings in Punjabi language. One word can be spelled more than one way.

ORIGIN AND SYMBOLS OF GURMUKHI SCRIPT

Gurmukhi is the name of the script used in writing primarily Punjabi and secondarily in the Sindhi language. The word Gurmukhi seems to have gained currency from the use of these letters to record the sayings coming from the mukh (lit. mouth or lips) of the Sikh Gurus [3].

It is commonly accepted that Gurmukhi is a member of the Brahmi family. Brahmi is an Aryan script which was developed by the Aryans and adapted to local needs.

As shown in the following figure 1.2, some of the major properties of Gurmukhi alphabets are:

- There are three vowel-carrier letters ੳ, ਅ and ੲ and nine vowel signs. The vowel-carrier ੳ and ਅ are never used without a vowel sign.
- There are 38 consonants in Gurmukhi alphabet.
- There are 3 half characters in Gurmukhi alphabet.
- In addition , there are other three signs bindi (ਂ), tippi (ਂ)and adhak (ੱ).

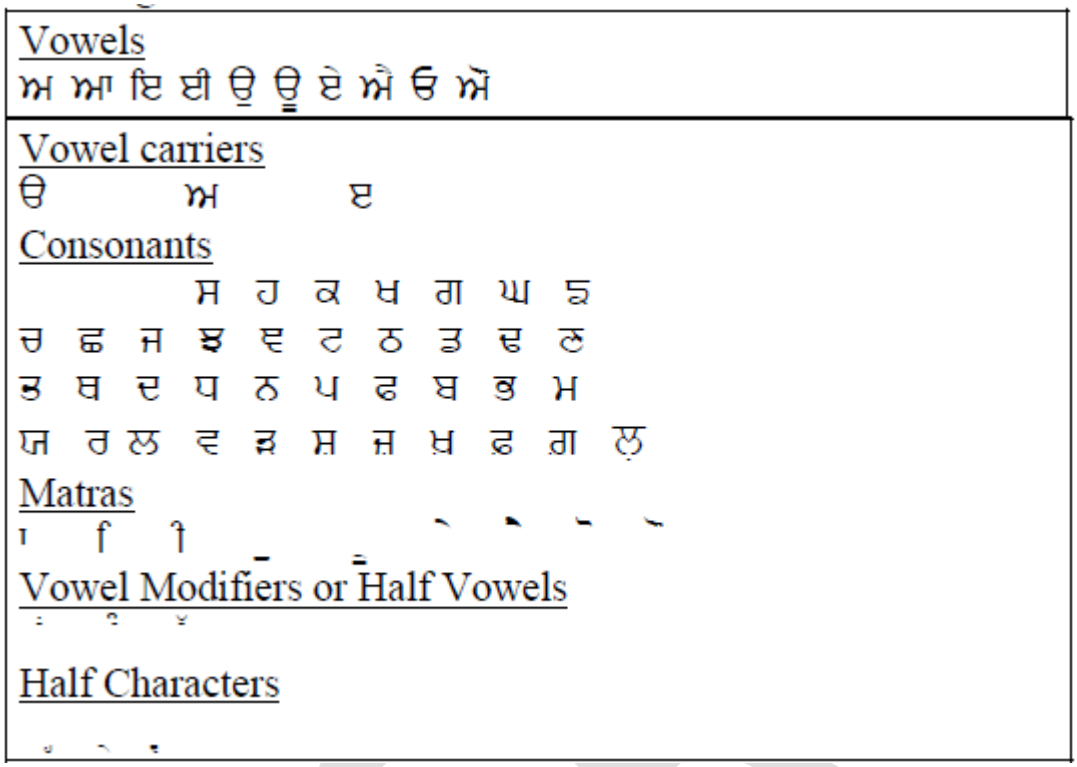

Figure 1.2 Gurmukhi Alphabet Set[1]

(a) ***Spell Checker Architecture***

The major components of a Punjabi spell checker are explained below as shown in the figure 1.1 above:

(b)*Tokenization*

We are assuming that the input text file will be in Unicode format. After the input, tokenization is the first step. Tokenization is the process to break the block of text into a list of words. The text is broken with the help of some boundary delimiter and blank spaces. The boundary delimiters here are several punctuation marks. The boundary delimiter could differ from font to font. In ASCII font several punctuation marks are part of the text but it is not the case in UNICODE encoding system. But there are certain punctuation delimiters which are to be ignored. One of them and the most important one is hyphen (-). Some words in the dictionary contains hyphen. Thus, hyphen cannot be used as a as a boundary delimiter.

(c) ***Error Detection Module***

The error detection module is the main component of the spell checker application. In this module each and every token is passed through a process through which the misspelled words are detected.

The technique used in the error detection module is **Dictionary Look-Up.** The tokens are searched for their presence in the dictionary. If the token is present in the dictionary, the word is correct. If the token is not found in the dictionary it means the word has been misspelled.

(d) ***Error Correction Module***

After the tokens have been through the error detection module, the errors are gone through the error correction module. The error correction module itself works in two phases, first is generation of a suggestion list and second is the ranking of the suggestion list as shown in figure 1.3.

(e) ***Generation of Suggestions***

For generation of the suggestion list, we have to have a proper approach which is both less complex and less time consuming. The most common approach used in the system is the minimum edit distance technique. In the minimum edit distance technique, the distance between the token and the dictionary word is calculated. A threshold value is decided according to which the words from the dictionary are considered as the candidate suggestions. If the distance between the token and the dictionary word is one then the word is considered as the candidate suggestion.

(f) ***Ranking of Suggestions***

As the suggestions has been generated, now it is very important to rank the suggestions according to an appropriate criteria which would bring the most suitable candidate word to the top of the list, so that the user don't have to search for the most appropriate suggestion throughout the suggestion list.

The suggestions are ranked keeping in mind two points:

- Phonetic similarity between the token and the candidate suggestion.
- The frequency of occurrence of the candidate suggestion.

The phonetic similarity is very important between the token and the candidate suggestion. The more the two words are phonetically similar, more the probability of that word to be the correct

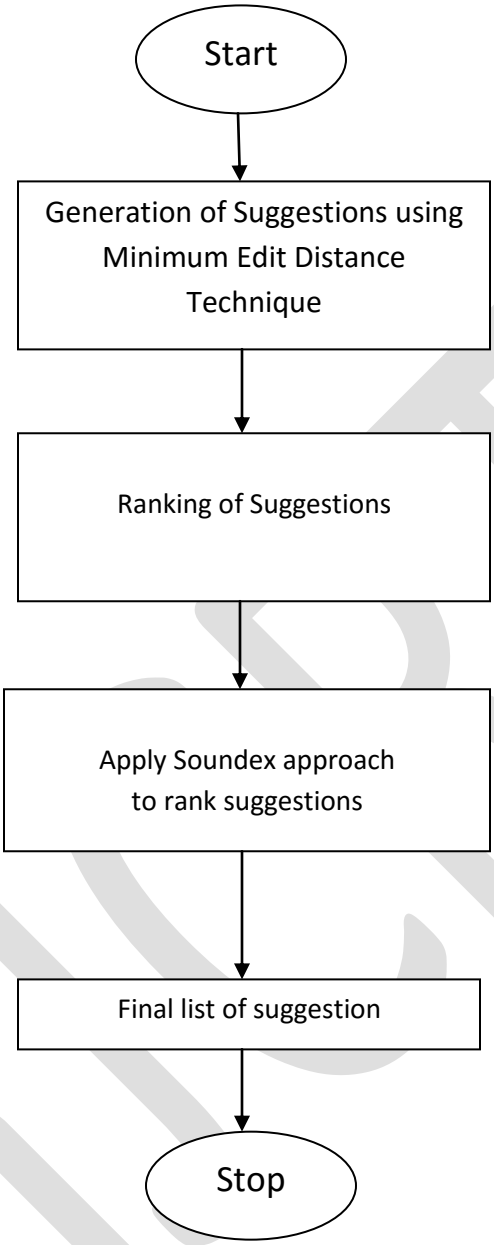

Figure 1.3 Flowchart representing the working of error correction module

spelling of the misspelled token. To find the phonetic similarity between two words the system uses an approach which is called the SoundexApproach.

(g) Soundex Approach

In the Soundex approach, a code value is applied to the similar sounding characters. Each suggestion is assigned a value with respect to the token. The value is assigned on the basis of the difference in characters of both the words where the characters are of similar sounds. For example, users often gets confused in characters like

ਗ-ਘ, ਜ-ਝ, ਦ-ਧ, ਬ-ਭ, ਡ-ਢ, ਨ-ਣ

ਸ਼-ਸ, ਖ-ਖ਼, ਗ-ਗ਼, ਜ਼-ਜ, ਫ-ਫ਼

ਿ-ੀ, ੇ-ੈ, ੋ-ੌ, ੁ-ੂ, ਂ-ਁ

These characters produce the similar sounds but are different and are the major cause of the misspelled words. Thus, using Soundex approach a similarity is generated between two words. Now the suggestions are assigned to the suggestion list and this list contains the most appropriate suggestions.

TEST WORDS

We have tested the system by giving the input from a well known Punjabi newspaper and some text form Punjabi stories from the internet where our system has given zero errors because the words of that text were already present in the dictionary. Then we intentionally made some errors in the same text and when that text of 90 words isgiven as an input then the system gave the correct suggestions for 80 words and there was no suggestion for 10 words i.e. those words were not present in the dictionary. It showed the correct word in suggestion list for 80 words out of 90 words that shows the accuracy of 90% for error correction. The table which contains the correct words, words in which errors are inserted, the Spell Checker result for error detection and correction is shown below in table 1.4.

TABLE 1.4 (Results)

Misspelled Word	Correct Word	Rank of the Correct Suggestion
ਕੱਠੇ	ਇਕੱਠੇ	1
ਫ਼ਿਰ	ਫਿਰ	1
ਉਪਰੋਥਲੀ	ਉੱਪਰੋਥਲੀ	1
ਉਭਰਿਆ	ਉੱਭਰਿਆ	1
ਅਹੁੱਦੇ	ਅਹੁਦੇ	Not Found
ਅਥਰੂ	ਅੱਥਰੂ	1
ਜਗਾਹ	ਜਗਾ	6
ਦ੍ਰਿੜਤਾ	ਦ੍ਰਿੜ੍ਹਤਾ	1
ਥੋਡੇ	ਥੋੜੇ	1
ਵਹੜੇ	ਵਿਹੜੇ	2
ਵੇਕਦੇ	ਵੇਖਦੇ	6
ਅਸਤੀਫਾ	ਅਸਤੀਫ਼ਾ	1
ਸੁਰੂਆਤ	ਸ਼ੁਰੂਆਤ	1
ਵਿਸ਼ਵਾਸ਼	ਵਿਸ਼ਵਾਸ	1
ਕ੍ਰਿਆ	ਕਰਿਆ	1
ਪਰਧਾਨਗੀ	ਪ੍ਰਧਾਨਗੀ	1
ਪ੍ਰੰਪਰਾਵਾਂ	ਪਰੰਪਰਾਵਾਂ	1
ਪ੍ਰਿਵਾਰ	ਪਰਿਵਾਰ	1
ਲਾਇਬਰੇਰੀ	ਲਾਇਬ੍ਰੇਰੀ	Correct Word
ਅਸਾਨੀ	ਆਸਾਨੀ	2
ਅਡੀਸ਼ਨਲ	ਐਡੀਸ਼ਨਲ	Not Found
ਅਥਾਰਟੀ	ਅਥਾਰਿਟੀ	1
ਅਧਾਰ	ਆਧਾਰ	2
Misspelled Word	**Correct Word**	**Rank of the Correct Suggestion**
ਅਧਿਅਨ	ਅਧਿਐਨ	1

ਆਜੇ	ਅਜੇ	1
ਸਬੰਧਤ	ਸਬੰਧਿਤ	1
ਸਾਬਿਤ	ਸਾਬਤ	3
ਸੈਨਟਰੀ	ਸੈਨੇਟਰੀ	1
ਹੋਇਗਾ	ਹੋਏਗਾ	1
ਕਾਲਿਜ	ਕਾਲਜ	1
ਖੜਾ	ਖੜ	2
ਖੰਿਡ	ਖਿੰਡ	1
ਗਿੱਦੜਬਹਾ	ਗਿੱਦੜਬਾਹਾ	1
ਜਹੀਆਂ	ਜਿਹੀਆਂ	1
ਬਲਾਉਂਦੇ	ਬੁਲਾਉਂਦੇ	Not Found
ਉਦਾਹਰਣ	ਉਦਾਹਰਨ	1
ਓਪਰਿਆ	ਓਪਰਿਆਂ	1
ਆਚਰਣ	ਆਚਰਨ	1
ਸਿਘ	ਸਿੰਘ	1
ਸੁਨਣ	ਸੁਣਨ	1
ਗਾਇਨ	ਗਾਇਣ	2
ਚੜ੍ਹਣ	ਚੜ੍ਹਨ	1
ਚੁਣੌਤੀਆਂ	ਚੁਨੌਤੀਆਂ	1
ਜਾਨਣ	ਜਾਣਨ	1
Misspelled Word	**Correct Word**	**Rank of the Correct Suggestion**
ਝੰਜੋੜਣ	ਝੰਜੋੜਨ	1
ਦਿੰਦਿਆ	ਦਿੰਦਿਆਂ	1
ਇਸਤੋਂ	ਇਸ ਤੋਂ	Not Found
ਹੈੱਡਕੁਆਰਟਰ	ਹੈੱਡ ਕੁਆਰਟਰ	1
ਏਅਰਲਾਈਨਸ	ਏਅਰਲਾਈਨਜ਼	1

ਗੰਮ	ਗ਼ਮ	Not Found
ਬਜੇ	ਵਜੇ	1
ਇੱਨਾ	ਇੰਨਾ	3
ਹਰਇਕ	ਹਰੇਕ	1
ਬੈਟਰੇ	ਬੈਟਰੀ	1
ਸਖਿਰ	ਸਿਖਰ	3
ਸਰਟੀਫਿਕਟੇ	ਸਰਟੀਫਿਕੇਟ	1
ਖੁੱਲ੍ਹਿਆ	ਖੁੱਲ੍ਹਿਆ	1
ਗਤਵਿਧੀਆਂ	ਗਤੀਵਿਧੀਆਂ	1
ਅਜ	ਅੱਜ	1
ਨਾਨ	ਨਾਨਕ	Correct Word
ਰਹਾ	ਰਿਹਾ	Correct Word
ਬਟਾ	ਬਿੱਟਾ	Correct Word
ਹਲਾਵਰਾਂ	ਹਮਲਾਵਰਾਂ	1
ਵਿਸਥਾ	ਵਿਵਸਥਾ	2
ਕਠੀਆਂ	ਕਾਠੀਆਂ	1
ਲੋਕ	ਲੋਕ	Correct Word
ਲੈਣਦੇਣ	ਲੈਣ-ਦੇਣ	Not Found
ਨੌਵਾਨ	ਨੌਜਵਾਨ	1
ਸਨੂੰ	ਸਾਨੂੰ	1
ਨੇੜ	ਨੇੜੇ	Correct Word
ਪ੍ਰਟਾਵਾ	ਪ੍ਰਗਟਾਵਾ	1
Misspelled Word	**Correct Word**	**Rank of the Correct Suggestion**
ਸਥਾਕ	ਸਥਾਨਕ	1
ਰਹ	ਰਹਾ	1
ਹੇ	ਰਹੇ	Correct Word
ਗਲੀਆਂ	ਗੋਲੀਆਂ	Correct Word
ਦਿੱਤ	ਦਿੱਤਾ	Correct Word
ਦਿਨਦਿਹਾੜੇ	ਦਿਨ-ਦਿਹਾੜੇ	Not Found
ਚਮਾ	ਚਕਮਾ	2
ਪਿਤੌਲ	ਪਿਸਤੌਲ	1
ਹਇਆ	ਹੋਇਆ	1
ਦੋਵ	ਦੇਵ	1
ਦਵੇਂ	ਦੋਵੇਂ	4
ਤਰ੍ਹਾ	ਤਰਾਂ	1
ਦਿੰਦਆਂ	ਦਿੰਦਿਆ	1

ਫਲ਼	ਫੇਲ਼	Not Found
ਜਾਕਾਰੀ	ਜਾਣਕਾਰੀ	1
ਤੇ	ਅਤੇ	Correct Word
ਕਾਯਾਬ	ਕਾਮਯਾਬ	1
ਅਨ	ਅਮਨ	Correct Word
ਗਲੀ	ਗੋਲੀ	1

CONCLUSION

This research work has accomplished the study of the Gurmukhi Script, the study of various spell checkers for Indian Languages and the techniques which are used to develop a spell checker by overcoming all the complexities which are present in Indian Languages. Also various types of error patterns are studied which occur in Punjabi text. In future, the system could enhance perhaps detect real word errors.

REFERENCES

[1] Gurpreet Singh Lehal, *Design and Implementation of Punjabi spell checker*, International journal of systemic cybernetics and informatics, 2007, pp.70-75.

[2] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3168328/

[3] http://www.omniglot.com/writing/punjabi.htm

[4] RupinderdeepKaur and Parteek Bhatia, *Design and Implementation of SUDHAAR-Punjabi Spell Checker*, International Journal of Information and Telecommunication Technology, Vol. 1, Issue 15 May, 2010, pp. 10-15.

[5] http://www.computerhope.com/jargon/s/spelchec.htm

[6] Bainu, *Online Spell Checker for Gurmukhi Script-SODHAK*, 2014p-7.